In silico prediction of Genomic Islands in microbial genomes

Georgios S. Vernikos
Selwyn College
University of Cambridge

April 2008

A dissertation submitted for
the degree of Doctor of Philosophy
at the University of Cambridge

Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute in Cambridge between April 2005 and April 2008. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this dissertation has been, or is being submitted for a degree or diploma or other qualification at this or any other university.

Georgios S. Vernikos
Cambridge
April 2008

Summary

Large inserts of horizontally acquired DNA that contain functionally related genes with limited phylogenetic distribution are often referred to as Genomic Islands (GIs). Integration of GIs can in a single step transform a microbial organism, changing radically the way it interacts with its niche, a process that is often referred to as "evolution in quantum leaps". A key aspect in the prediction of GIs is that at the time of their integration in the new host chromosome, their composition reflects mainly the composition of the donor, rather than the host.

The first part of this thesis concerns the design, development and implementation of a novel algorithm, that of the Interpolated Variable Order Motifs, for the composition-based prediction of GIs. This algorithm exploits compositional biases using variable order motif distributions and captures more reliably the local composition of a sequence compared with fixed order methods, overcoming the limitations of the latter. Furthermore, for the optimal localization of the boundaries of each predicted region, the Hidden Markov Model theory was implemented in a change point detection framework, predicting more accurately the true insertion point of candidate GIs.

In the second part of this thesis whole genome based comparative and phylogenetic techniques were used to study the acquisition of horizontally acquired genes in the *Salmonella* lineage in a time dependent manner. The compositional amelioration process was modelled and the relative time of acquisition of those genes was determined on different branches of the *S. enterica* phylogenetic tree.

The aim of the third part of this thesis is to explicitly quantify and model the contribution of genomic features to the GI structure, under a probabilistic framework. A hypothesis free, bottom-up search was implemented and identified approximately 700 genomic regions, including

both GIs and randomly sampled regions from three different genera that form my training dataset. A Machine Learning approach was used to exploit the above dataset and study the structural variation of GIs.

The last part of this thesis focuses on the experimental validation of the *in silico* predictions made on a newly sequenced bacterial genome. Applying a PCR-based protocol, the presence and absence of the predicted candidate islands was probed in seventeen unsequenced closely and distantly related strains. The true borders of the predicted islands were confirmed by sequencing across the boundary site in strains lacking the island.

Acknowledgements

Firstly, my deepest personal thanks to my supervisor Julian Parkhill for being the busiest and at the same time the most available "teacher" that shaped, in a very profound and enjoyable way, my academic career; Julian has a remarkable ability to inspire people working next to him.

I would also like to thank members of my thesis committee, George Salmond for discussion on various biological aspects of my project, Richard Durbin and Alex Bateman for ideas and discussion on algorithmic techniques and Thomas Down for replying to my emails using pseudo-code! – It made our communication much easier and fun.

Many thanks to David Carter for making the source code of the biojava implementation of the Relevance Vector Machine available, Nick Thomson for long, late night philosophical (and not only) discussions on microbial genomes; Stephen Bentley for biological discussion and for contributing to a key step of my future career path, Matthew Holden for discussion on various aspects of microbial evolutionary dynamics and the other members of team 81 for advice on different steps of this project.

I also thank: Helena Seth-Smith and Paul Scott, for being extremely patient with my ignorance on lab-based techniques and Andrew Jackson for discussion on aspects of the phylogenetic analysis; Xavier Didelot for advice on whole genome sequence alignments and Matthew Avison for providing the *Stenotrophomonas maltophilia* strains; Manolis Dermitzakis for my very first, off-the-record, coffee-interview that took place in Greece; the Wellcome Trust Sanger Institute for my PhD studentship.

Special thanks to my parents, Stelios and Maria, for letting me take apart their washing machine, in order for me to discover what lies beneath; their very liberal approach to my intellectual curiosity shaped in a very profound way my academic and non-academic way of thinking; my brother, Christos, for convincing me that different is nice.

My love to Andriana (my wife to be) for standing by me no matter what, without whom at the end of this 4-year academic journey, I would only be a scientist.

Contents